Genomics 100 (2012) 57–63

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

AncestrySNPminer: A bioinformatics tool to retrieve and develop ancestry informative SNP panels

Sushil Amirisetty, Gurjit K. Khurana Hershey, Tesfaye M. Baye *

Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati, Cincinnati, OH, USA

ARTICLE INFO

Article history:
Received 12 January 2012
Accepted 6 May 2012
Available online 11 May 2012

Keywords:
Ancestry
Ancestry informative markers
AIMs
Bioinformatics
AncestrySNPminer
Data mining
Admixture
Admixture mapping

ABSTRACT

A wealth of genomic information is available in public and private databases. However, this information is underutilized for uncovering population specific and functionally relevant markers underlying complex human traits. Given the huge amount of SNP data available from the annotation of human genetic variation, data mining is a faster and cost effective approach for investigating the number of SNPs that are informative for ancestry. In this study, we present AncestrySNPminer, the first web-based bioinformatics tool specifically designed to retrieve Ancestry Informative Markers (AIMs) from genomic data sets and link these informative markers to genes and ontological annotation classes. The tool includes an automated and simple "scripting at the click of a button" functionality that enables researchers to perform various population genomics statistical analyses methods with user friendly querying and filtering of data sets across various populations through a single web interface. AncestrySNPminer can be freely accessed at https://research.cchmc.org/mershalab/AncestrySNPminer/login.php.

1. Introduction

High-throughput genomic technologies are generating dense, rich, high quality measures of human genomic variations in individuals and across the genome simultaneously [1]. Much of these data are deposited either at the institution-based supercomputing (or cloud) storage facilities or at the national databases center that are openly available to the research community to enable researchers utilize the data to develop new information and knowledge. Methodological advances in the analysis and filtering relevant information from these data have simply not kept pace with this flood of data [2]. The utility of our monumental investment in data generation will ultimately depend on our ability to extract the maximal amount of information from these genomic data and enhance our ability to gain knowledge about the genetic architecture of human genomic variation and disease genetics. Currently, only a limited amount of information characterizing SNPs across the human genome for major ethnic groups is found in the literature [3–5]. Consequently, mining of informative SNP markers from such high genomic resolution data sets is an economical, rapid, and practical strategy for developing a more comprehensive and informative panel for ancestry [6,7]. This may result in a uniform resource that describes nucleotide diversity with sufficient power to infer ancestry for various populations [8].

When inferring genetic ancestries, investigators have found that some markers (or variants) are more informative than others [9,10] and most often alleles common in one population are common in other populations [3,11]. As a result, several measures of marker informativeness (ability of markers to differentiate between populations) have been developed to select the most informative ancestry informative markers (AIMs) from an ever-increasing wealth of genomic databases [12]. These measures include absolute allele frequency differences (Δ), Shannon information content (SIC), Fisher information content (FIC), F statistics index (F_{ST}), and the informativeness for assignment measure (I_n) [12]. These measures of informativeness to select AIM panels along with a new composite measure (CompM) were recently computed and compared by Ding et al. [13]. CompM is an approach that combines and ranks informative markers from all the five measures [Δ, SIC, FIC, F_{ST}, and I_n]. The CompM rank was based on assigning a score for each marker and rank based on the average score of all the measures. Hence, using CompM, SNPs can be ranked according to their allelic frequency differences between ancestral populations, and the top-ranked SNPs based on the average rank score using all the five measures will be used to develop comprehensive and informative AIMs panel. We showed that this combined measure of ranking outperformed all the five measures when used individually [13].

Currently more than 38 million common and rare variants have been deposited in public databases (dbSNP, Build 132), and users of these data sets face significant analytical challenges, including how to easily access these databases in real-time and determine which of the 38 million polymorphisms are most informative to infer ancestry. Zubritsky [14] described the process analogous to gold mining, stating that prospectors

* Corresponding author at: Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, OH 45229, USA. Fax: +1 513 6361657.
E-mail addresses: sushil.amirisetty@cchmc.org (S. Amirisetty), gurjit.hershey@cchmc.org (G.K.K. Hershey), tesfaye.mersha@cchmc.org (T.M. Baye).

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2012.05.003

will invariably find pyrite mixed with gold. To handle such challenges there is a need to select the most informative SNPs and reduce the likelihood of evaluating monomorphic SNPs in any given data set. Removal of such uninformative SNPs reduces the time, genotyping costs and the noise while increasing the power for identification/extraction of the most informative panels of SNPs [15]. Consequently, an in silico method of data mining is needed to find and extract non-redundant AIMs from the ever-increasing wealth of genomic data in both private and public domains [16].

In this study, we examined genome-wide SNP data among human populations using Δ, SIC, FIC, F_{ST}, I_n and CompM. To achieve this, we developed AncestrySNPminer, a web-based application that is capable of mining millions of SNPs automatically across different populations. AncestrySNPminer automates the tasks of extracting informative SNPs derived from different ancestral populations and provides several flexible ways to present and export the data. AncestrySNPminer is oriented towards flexible, user-defined SNP selection criteria such as chromosome number, chromosomal position, ancestral population, and measures of marker informativeness as well as spacing between markers. The application is publicly available, as a web based tool, at https://research.cchmc.org/mershalab/AncestrySNPminer/login.php. Changes in the evolving databases will be incorporated to improve its utility and generate updated data sets of interest.

2. Materials and methods

2.1. Data input

Publicly available data such as The International haplotype map (HapMap) and Human Genome Diversity Project (HGDP) as well as user generated data are considered as the data input. As the default analysis, HapMap (http://hapmap.ncbi.nlm.nih.gov/) Phase III data were used. A total of 1,397 samples from 11 populations and 3.2 million SNP markers were included. The 11 populations represent Africa, Europe, Asia and North America (Fig. 1).

The protocol used to develop AncestrySNPminer is the Common Gateway Interface (http://c2.com/cgi/wiki?CommonGatewayInterface). The Common Gateway Interface (CGI) is a standard protocol that defines how web server software can delegate the generation of web pages to a stand-alone application or an executable file. CGI scripts were written in Python programming language. Python provides a portable, interpreted, interactive, object-oriented programming (OOP) language ideal for Common Gateway Interface (CGI) Web programming [17]. Furthermore, the combination of significant power and clear syntax makes Python an excellent instructional tool. Language features include modules, classes, high-level dynamic data types, and dynamic typing.

When a user clicks the submit button, CGI script is invoked by an HTTP server to process user inputs submitted through an HTML <FORM> element. CGI script lives in the server's special cgi-bin directory. The HTTP server places all information about the job in the script's shell environment, executes the script, and sends the output back to the client's browser (Fig. 2).

2.2. AncestrySNPminer workflow

Fig. 3 illustrates a workflow-chart for an in silico analysis to select SNPs based on default human population and SNP database. The computer program written in Python is a dynamic query form that accesses the server using HTTP connection over the Internet. The webserver software executes the CGI script. The CGI script connects to the HapMap FTP and local database and executes many sql commands to retrieve the data. The webserver now sends the browser the HTML prepared by the CGI script. The final results of the query are organized in a tabular format, which can be reviewed by users (Fig. 3). AncestrySNPminer relies on a local relational database and a real-time access to the ancestral populations and SNP data deposited such as HapMap or, HGDP-CEPH FTP. A local database is created by mining biological information from NCBI, Ensembl BioMart and HapMap databases. The result is a database containing about 38 million SNPs that map over 25,000 genes. The application is oriented toward flexible user-defined SNP selection criteria including either one or the combination of the following: Δ, SIC, FIC, F_{ST}, I_n and CompM. It provides several flexible ways to present and export the data.

2.2.1. Availability and requirements

Project name: AncestrySNPminer

Project home page: https://research.cchmc.org/mershalab/AncestrySNPminer/login.php

Operating system: platform independent

Programming language: Python, MySQL, JavaScript, PHP, HTML, CSS

Programming language required by the user: None

Other requirements: JavaScript enabled web browsers

License: The tool is available online free of charge

Any restrictions to use by non-academics: None

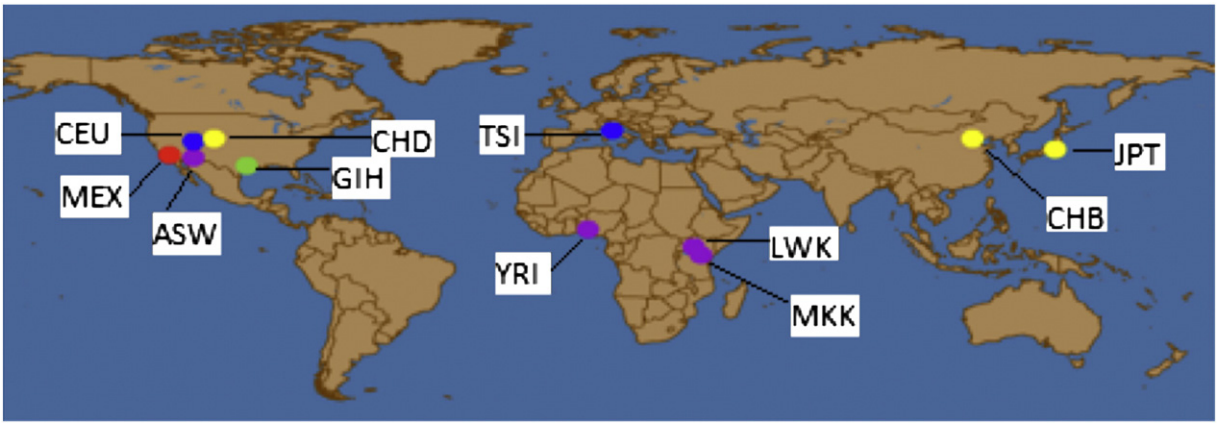

Fig. 1. Geographic map of the HapMap phase III world populations. CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; TSI, Toscani in Italia, JPT, Japanese in Tokyo, Japan; CHB, Han Chinese in Beijing, China; CHD, Chinese in Metropolitan Denver, Colorado; GIH, Gujarati Indians in Houston, Texas; MXL, Mexican ancestry in Los Angeles, California; LWK, Luhya in Webuye, Kenya; ASW, African ancestry in Southwest USA; MKK, Maasai in Kinyawa, Kenya; YRI, Yoruba in Ibadan, Nigeria. The distribution of allele frequencies across the genome and various ancestral populations were extensively investigated by using several measures of marker informativeness and the HapMap database. Over 4 million SNPs were compared between HapMap populations.

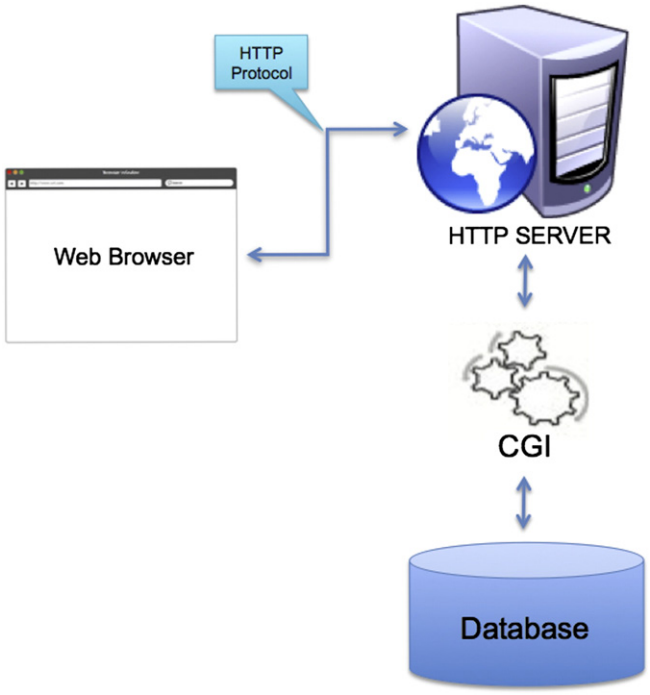

Fig. 2. Programmatic architecture of AncestrySNPminer consists of a CGI program placed in cgi-bin server. It is developed using Python scripts that connect to relational database (MySQL) with gene and gene associated ontology information and HapMap FTP site. A user interface was then developed using HTML, Java script.

3. Results

3.1. Programmatic architecture

AncestrySNPminer web interface runs using any modern web browser such as Firefox, Internet Explorer, Chrome, Safari or Opera. The web application was developed using XHTML, JavaScript and CSS at the client-side, Python, PHP at the server-side, and MySQL as a database. AncestrySNPminer architecture has two layers: an access layer with a web server and a storage layer with a MySQL database server. Python annotation pipeline was used to connect the user data to the source database and perform all the measures of marker informativeness calculations. Primary data sources for the attributes are Ensembl BioMart, NCBI gene, and dbSNP FTP. Gene information and gene ontology information data are downloaded from NCBI gene and GO FTP site. Gene associated information is downloaded from Ensembl BioMart. SNP data are derived from HapMap FTP and dbSNP.

3.2. Components of AncestrySNPminer

The window displaying the AncestrySNPminer web application interface is shown in Fig. 4, which has three key distinct components. The top layer of the window contains links for manual, frequently asked questions, and feedback followed by an option to select a database (either user owned, HapMap, HGDP or 1000 genomes project). Currently, user generated, HapMap and HGDP databases are supported. In the future as the data become more stable, the 1000 genomes project data sets will be included.

The next component includes a text area where users may enter a list of SNPs or genes to find out the informativeness measure between any two populations, or enter the start and end positions of the chromosome for a chromosome query to investigate between population measures of marker informativeness in specific genomic region of interest. In both cases, user must select populations under Population 1 and Population 2. The measures for marker informativeness to choose include Δ, FIC, SIC, F_{ST}, I_n and CompM. Filters can be selected by clicking on the radio buttons or by entering values manually in the space provided. Spacing between markers can be specified to set a physical distance in order to decrease the chance that selected AIMs are in strong LD. Users can select a distance ranging from 50 kb to 5 Mb or can enter a custom value. This query is optional. In the bottom layer of the application window, the user can select the result to be displayed on the browser by clicking the Display button or select that the file be downloaded in either tab delimited text or Excel format (Fig. 4).

3.2.1. Optional attributes

In addition to the main application, which is to prioritize markers for ancestry informativeness, AncestrySNPminer allows the user (optionally) to further mine different attributes for the selected AIMs panel under the categories gene information, gene associated information and gene ontology information. These optional attributes are listed on the left side of the application. They are divided into three categories; gene information includes gene symbols and descriptions, and gene associated information includes Ensembl gene ID, SNP classification/category based on genomic position (e.g., exon, promoter, intron etc.) or based on predicted functional effect (e.g., non-synonymous, synonymous, etc.) and sequence variation. Gene ontology information includes molecular function, biological process and cellular component of the gene that have been mapped to selected variants. SNP "chips" from various genotyping platforms such as Affymetrix 6.0 and Illumina 1 M arrays were included as well. All this information is available to the user in a single location

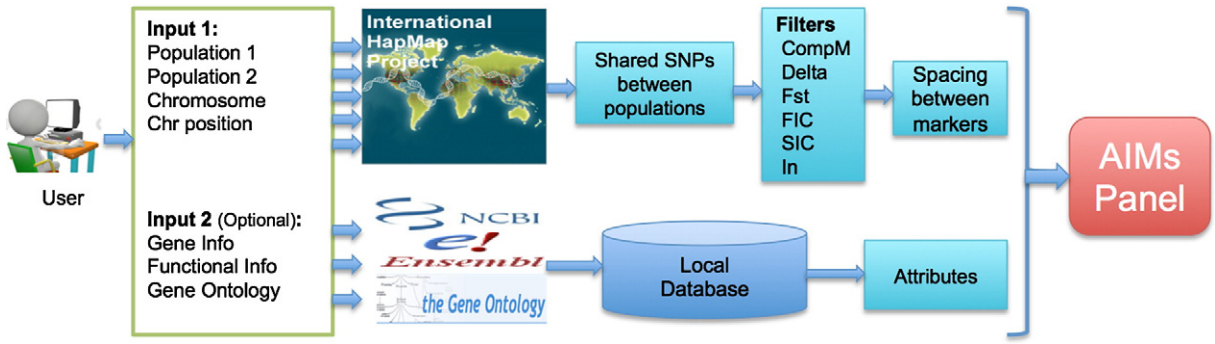

Fig. 3. AncestrySNPminer workflow strategies. As shown in the flow diagram, once the user enters the information and clicks the submit button, information is mined from HapMap FTP and local databases, and AncestrySNPminer algorithm computes all pairwise comparisons for all measures of marker informativeness to output Ancestry Informative Markers (AIMs).

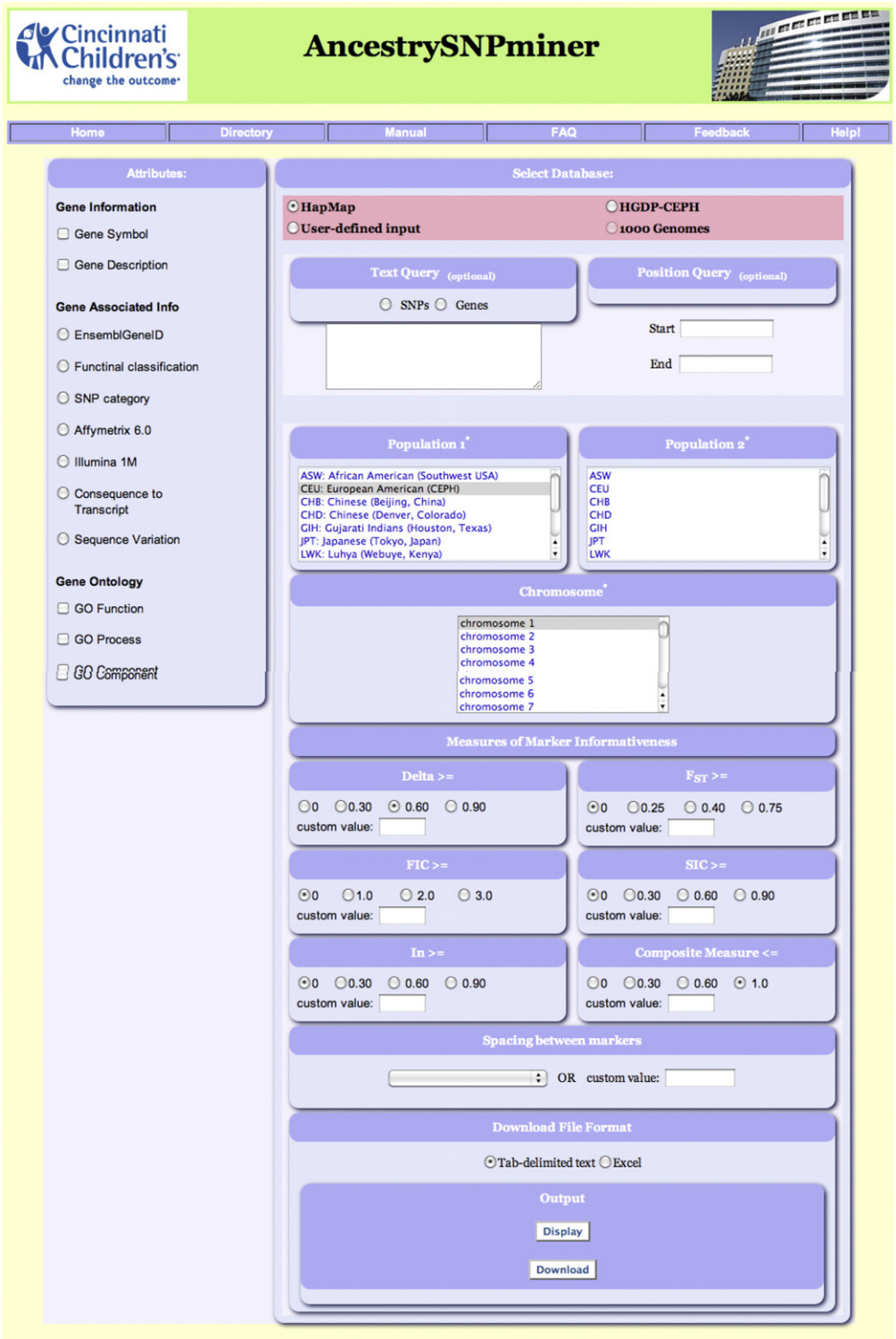

Fig. 4. AncestrySNPminer web interface. It is designed as an architecture consisting of three layers where a user selects populations, chromosomes and various filters including measures of marker informativeness to generate AIMs. The web interface allows user to either display the information on the browser or to download to a local hard drive.

under a unified interface. Users can use AncestrySNPminer to identify SNPs with major ancestry information and link with genes and chromosome region by mining bioinformatics databases from the National Center for Biotechnology Information Entrez Gene, Ensembl, OMIM, etc. to objectively identify the relevance of the variant in relation to population differentiation. For example, in order to understand the biology of highly differentiated SNPs based upon annotated functional class, one can assign SNPs to either genic or non-genic and further subdivided genic SNPs into either synonymous or non-synonymous categories (all non-genic SNPs were categorized as synonymous). We can link the variant to genes and gene products to determine: (a) their known or predicted molecular function (e.g. type of biochemical activity), (b) cellular locale (e.g. nucleus), or (c) their biological role (e.g. transcription) (www.geneontology.org).

3.2.2. Data output

3.2.2.1. Data export. Data can be exported as a tab-delimited format or excel format. Output files include query details including AIM panels with assigned rs numbers, allele frequencies in each population selected, chromosome numbers, Δ, F_{ST}, FIC, SIC, I_n, CompM and the chromosome positions. All marker information generated using this tool is freely and publicly available to non-commercial users without restriction.

3.2.2.2. Data visualization. Users can view the data on the web browser by clicking the Display button after selecting the parameters. The output file is generated through HTML and displayed in a new window of the web browser. All contents are displayed in a tabular format (Table 1).

3.2.3. Example

To assess AncestrySNPminer's applicability in a data mining workflow, case studies of queries that can be used to describe a set of SNPs were outlined as "retrieve all SNPs in the first million bases of chromosome 4, 7 and 8 between any two populations along with attributes such as gene information, gene associated information and gene ontology terms; retrieve the SNPs belonging to all the chromosomes from position 10,000 to 10,000,000 between any two populations with attributes"; "Type in or paste Genes or SNP IDs in the text area separated by a new line and query by selecting either one filter or multiple filters at a time between two populations"; "mine SNP's between any two populations using filters and spacing to account linkage disequilibrium of 100 kb." Additional information displayed for each SNP includes the rsID, chromosome number, minor allele, and allele frequencies in both populations, as well as the Δ, F_{ST}, FIC, SIC, I_n and CompM values and attributes. All of this information is available to the user in a single location and under a unified interface. Using Pritchard et al.'s [18] STRUCTURE computer program and the top 500 ancestry informative SNPs that we generated using compM and real samples/data, we show how AncestrySNPminer can successfully cluster West African (YRI, $n=203$), African American (ASW, $n=87$) and European (CEU, $n=164$) ancestry individuals into their respective geographic population membership (Fig. 5) and estimate the percent

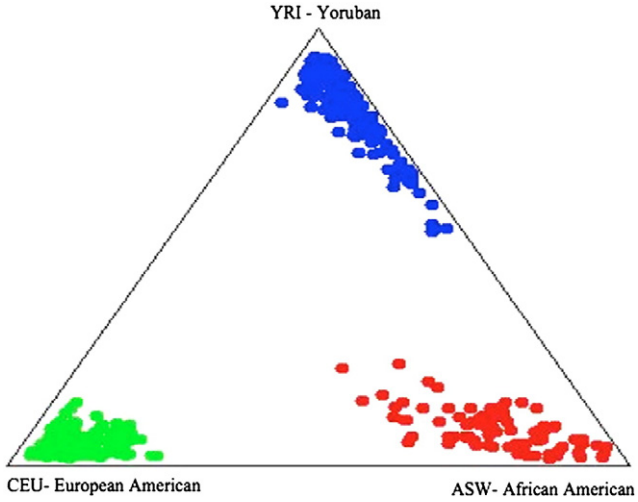

Fig. 5. Triangle clustering plots of parental populations. Each individual is represented by a colored point. The colors correspond to the prior population labels. Each point shows the mean estimated ancestry for an individual in the sample. For a given individual, the values of the three coefficients in the ancestry vector q are given by the distances to each of the three sides of the equilateral triangle. After the clustering was performed, the points were labeled according to sampling location. The estimated ancestry vector for an individual consists of K components which add up to 1. At $K=3$, the ancestry vectors are plotted onto a triangle, as shown. In these plots, the closer a person is to one of the three-labeled axis, the greater his or her ancestry proportion is from that population. Individuals who are in one of the corners are therefore assigned completely to one population or another.

Table 1
Output displayed in the browser when a user clicks the display button. This is an example of partial view for CEU versus YRI in chromosomes 2 and 3 using filters $\Delta>=0.6$, $F_{ST}>=0.25$, FIC$>=1.0$, SIC$>=0.3$ and $I_n>=0.3$ and CompM$<=1.00$.

Delta file of CEU - YRI; Chromosome(s): ['1', '4', '6'];

Input values : Delta >= 0.600000 Fst >= 0.250000 FIC >= 1.000000 SIC >= 0.300000 In >= 0.300000 CompM <= 1.000000
Start Position > 11668926 End Position < 253000000

rsID	chr	pos	Allele	CEU	YRI	delta	Fst	FIC	SIC	In	CompM
rs6674304	chr1	116689265	T	0.92	0.02	0.90	0.82	5.03	0.36	0.73	0.022727
rs12087334	chr1	116688978	C	0.92	0.02	0.89	0.80	4.96	0.35	0.71	0.068182
rs9306906	chr4	33642762	C	0.88	0.01	0.87	0.76	5.03	0.35	0.68	0.159091
rs4839518	chr1	116547474	G	0.11	0.98	0.87	0.76	4.91	0.34	0.68	0.181818
rs9321552	chr6	136523305	G	0.99	0.08	0.91	0.84	4.29	0.34	0.77	0.204545
rs1827950	chr4	117317931	G	0.11	0.97	0.86	0.76	4.82	0.34	0.67	0.227273
rs3823159	chr6	136524420	A	0.99	0.08	0.91	0.83	4.28	0.34	0.76	0.227273
rs6446975	chr4	75254908	A	0.06	0.95	0.88	0.78	4.50	0.33	0.68	0.250000
rs2759281	chr1	203132988	T	0.86	0.01	0.85	0.74	4.94	0.34	0.66	0.272727
rs7753890	chr6	136557950	T	0.98	0.08	0.90	0.82	4.22	0.33	0.74	0.272727
rs1448275	chr4	32726082	T	0.11	0.97	0.86	0.74	4.76	0.33	0.65	0.318182
rs3734548	chr6	136550092	T	0.98	0.08	0.89	0.80	4.12	0.32	0.72	0.318182
rs719776	chr4	33363055	G	0.88	0.02	0.86	0.74	4.70	0.33	0.65	0.340909
rs835574	chr1	120264753	C	0.87	0.01	0.85	0.73	4.77	0.33	0.65	0.363636
rs11264110	chr1	35408814	G	0.09	0.96	0.86	0.75	4.50	0.32	0.65	0.409091

*Δ = absolute allele frequency differences; SIC = Shannon information content; FIC = Fisher information content; F_{ST} = F statistics index; I_n = informativeness for assignment measure; CompM = composite measure.

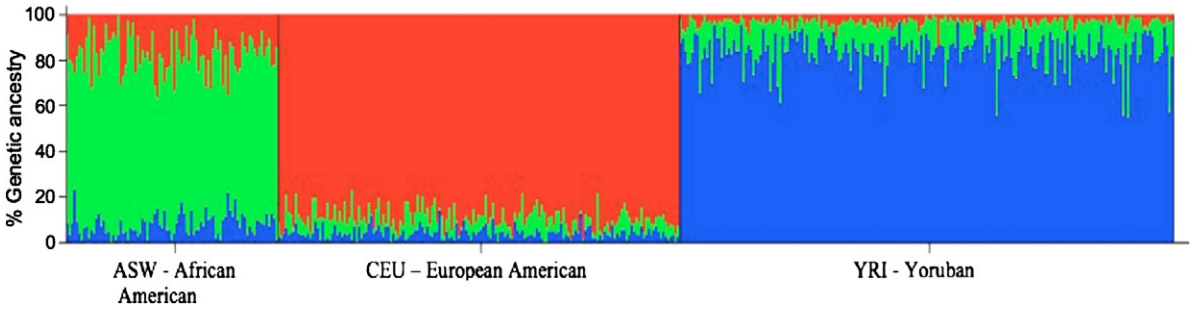

Fig. 6. Estimated population structure using model based clustering algorithm. Each individual is represented by a thin vertical line, which is partitioned into K colored segments that represent the individual's estimated membership fractions in K clusters. Black lines separate individuals of different populations. Populations are labeled below the figure, with their regional affiliations above it. Individuals can have membership in multiple clusters, with membership coefficients summing to 1 across clusters. There are clear differences in the individual admixture estimates for these three population samples. The parental populations are very much separated from each other and are localized into tight groups. The African American populations are much more varied in the individual ancestry levels that are observed within the two ancestral groups.

ancestry of each individual (Fig. 6). Although individuals were classified according to geographic origin using cluster analysis, we observed substantial variability in the ancestral genetic background. It is interesting that the continental AIMs show an ability to resolve among geographic population groups. This observation has additional implications for the utility of the panel in helping resolve elements of admixed population structure in human study samples. For instance, individuals from African American admixed populations showed intermediate coordinates between European and African ancestry. The ethnicity specific clustering within continental populations using continental AIMs thus indicates that these markers contain some ethnogeographical information, with a power to resolve among sub-continental populations with differing population histories. Besides the actual program, a step-by-step user manual has been developed and provided to help users and can be downloaded from the AncestrySNPminer website (Supplemental File 1). We have also listed answers for commonly asked questions in the Frequently Asked Questions (FAQ) tab.

3.2.4. Execution time

Execution time depends upon the number of filters including the number of chromosomes selected between two populations as well as the level of marker informativeness needed. Mining the information for all twenty-four chromosomes takes longer than mining information for two or three chromosomes. The filters also influence the execution time of the application. Estimated execution time for all the chromosomes between CEU-YRI populations without any filters and attributes resulting in more than 1.5 million SNPs takes 4 minutes. For example, the query between two populations (CEU and YRI) of chromosomes 2 and 4 gives an output of 185,908 SNPs. The same query using filters such as $\Delta > 0.7$ and $I_n > 0.3$ gives an output of 870 SNPs which takes less time to process.

3.3. How to use AncestrySNPminer?

The application is publicly available, as a web-based tool, at https://research.cchmc.org/mershalab/AncestrySNPminer/login.php website. The codes are available free of charge for all major platforms, and can be obtained upon request. The user needs to register in order to access AncestrySNPminer by filling out a simple registration form. Then, the users can login with their email ID. We welcome feedback from the user community. Changes in the evolving databases will be incorporated into AncestrySNPminer to improve its utility and to insure that it is up to date.

3.4. Future development

AncestrySNPminer will play an important role in ancestry inference, evolutionary genetics, and forensic genetics and population genetic structure studies. Future iterations of AncestrySNPminer will include data mining capability from the 1000 Genomes Project database as well as adding more mining options and biological information to the query form. For example, as genotype databases become more integrated with electronic medical records, we plan to link variation and mutation data to population specific clinical information in order to be able to interpret diagnostic information and serve patients and their relatives optimally.

4. Conclusion

In this study, we presented a novel genomic data-querying tool, AncestrySNPminer. AncestrySNPminer is designed to help users with limited bioinformatics skills to take advantage of their own or publicly available genomic data to mine and identify AIMs between ancestral populations by using one or multiple measures of marker informativeness. AncestrySNPminer provides several options (either specified genomic regions or genome-wide) to retrieve and link informative markers along with gene ontology and functional information. AncestrySNPminer will greatly enhance in silico data mining and analysis by allowing a fast and user friendly comparison of populations from user generated or publicly available genome-wide data, thus facilitating the transition from a single marker search to genome-wide queries in a cost-effective way. Our tool has several advantages: (1) the user does not need to access many web sites for multiple data sources. We offered a “one-stop shop” solution to ease access and management of a large variety of biological data from different data sources. (2) All queries requested by users are executed in real time. There are some limitations to be noted, for example, AncestrySNPminer is directly appropriate to develop ancestry informative markers from two ancestral populations at a time and it does not account for multi-allelic situations at a locus. However, for admixed descendent from three ancestral populations such as Latino, one can still use AncestrySNPminer. Since the Latinos are admixed population from Africans, Europeans and Native Americans, the first step is to generate AIMs from pairwise genome-wide data between African and European, African and Native Americans, between European and Native Americans using AncestrySNPminer. The AIMs panel with low heterogeneity within continent should be given priority as they can be portable in Latino populations throughout the Americas. Validating the panel of AIMs using independent Latino populations and genome-wide data is recommended before investing millions of dollars on an admixture or ancestry related projects.

Finally, although primarily designed to leverage private and public databases to identify ancestry informative markers, AncestrySNPminer has a query option, which can be used in population based-study to identify genomic regions that show high genomic differentiation and natural selection between populations. This feature may help prioritize SNPs for selective genotyping (i.e. to prioritize or select SNPs for further individual genotyping and re-sequencing) as shown recently [19]. SNPs can be ranked according to their allelic frequency differences between

cases and controls, and the top-ranked SNPs and associated genes can be selected for further genotyping in replication studies as well as for further re-sequencing. The approach can also be used to develop matched case–control study design based on similarity in allele frequencies within groups [20]. Empirical assessment of differences in allele frequencies in both population-based or case–control studies taking biogeographical ancestry as covariate in the logistic regression model can shed light on the genetic architecture of diseases as well as potential for generalizability of findings across population groups [21]. AncestrySNPminer based-ancestral classification could be also used to assist in identity testing and inference of ancestry in a forensics setting. In summary, the easy to use, HTML display summary report, and the Excel or Tab-delimited download formatted AIMs reports with various filtering criteria allow the researchers to navigate, filter, and elucidate tens of thousands of variants to focus on potentially relevant ancestry informative markers in population genomic studies.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.05.003.

Acknowledgments

We thank T. Abebe for helpful discussion and comments. This work was supported by National Institutes of Health grant 1K01HL103165 and P30HL10133.

Web Resources

HapMap: http://www.hapmap.org/
HGDP: http://hagsc.org/hgdp/files.html
Python: http://www.python.org/
MySQL: http://www.mysql.com/
JavaScript: http://www.javascript.com/
PHP: http://www.php.net/
NCBI: http://www.ncbi.nlm.nih.gov/
Ensembl: http://useast.ensembl.org/
Gene Ontology: www.geneontology.org
CGI: http://c2.com/cgi/wiki?CommonGatewayInterface

References

[1] L.L. Cavalli-Sforza, M.W. Feldman, The application of molecular genetic approaches to the study of human evolution, Nat. Genet. 33 (2003) 266–275 (Suppl.).
[2] T.M. Baye, Inter-chromosomal variation in the pattern of human population genetic structure, Hum. Genomics 5 (2011) 220–240.
[3] T.M. Baye, R.A. Wilke, M. Olivier, Genomic and geographic distribution of private SNPs and pathways in human populations, Per Med 6 (2009) 623–641.
[4] M.W. Smith, N. Patterson, J.A. Lautenberger, A.L. Truelove, G.J. McDonald, A. Waliszewska, B.D. Kessing, M.J. Malasky, C. Scafe, E. Le, P.L. De Jager, A.A. Mignault, Z. Yi, G. De The, M. Essex, J.-L. Sankale, J.H. Moore, K. Poku, J.P. Phair, J.J. Goedert, D. Vlahov, S.M. Williams, S.A. Tishkoff, C.A. Winkler, F.M. De La Vega, T. Woodage, J.J. Sninsky, D.A. Hafler, D. Altshuler, D.A. Gilbert, S.J. O'Brien, D. Reich, A high-density admixture map for disease gene discovery in african americans, Am. J. Hum. Genet. 74 (2004) 1001–1013.
[5] H.E. Collins-Schramm, C.M. Phillips, D.J. Operario, J.S. Lee, J.L. Weber, R.L. Hanson, W.C. Knowler, R. Cooper, H. Li, M.F. Seldin, Ethnic-difference markers for use in mapping by admixture linkage disequilibrium, Am. J. Hum. Genet. 70 (2002) 737–750.
[6] M.D. Shriver, M.W. Smith, L. Jin, A. Marcini, J.M. Akey, R. Deka, R.E. Ferrell, Ethnic-affiliation estimation by use of population-specific DNA markers, Am. J. Hum. Genet. 60 (1997) 957–964.
[7] X. Mao, A.W. Bigham, R. Mei, G. Gutierrez, K.M. Weiss, T.D. Brutsaert, F. Leon-Velarde, L.G. Moore, E. Vargas, P.M. McKeigue, M.D. Shriver, E.J. Parra, A genomewide admixture mapping panel for Hispanic/Latino populations, Am. J. Hum. Genet. 80 (2007) 1171–1178.
[8] A.L. Price, N. Patterson, F. Yu, D.R. Cox, A. Waliszewska, G.J. McDonald, A. Tandon, C. Schirmer, J. Neubauer, G. Bedoya, C. Duque, A. Villegas, M.C. Bortolini, F.M. Salzano, C. Gallo, G. Mazzotti, M. Tello-Ruiz, L. Riba, C.A. Aguilar-Salinas, S. Canizales-Quinteros, M. Menjivar, W. Klitz, B. Henderson, C.A. Haiman, C. Winkler, T. Tusie-Luna, A. Ruiz-Linares, D. Reich, A genomewide admixture map for Latino populations, Am. J. Hum. Genet. 80 (2007) 1024–1036.
[9] T.M. Baye, H. He, L. Ding, B.G. Kurowski, X. Zhang, L.J. Martin, Population structure analysis using rare and common functional variants, BMC Proc 5 (2011) S8.
[10] M.J. MacKinnon, N. Glick, Data mining and knowledge discovery in databases—an overview, Aust. NZ J. Stat. 41 (1999) 255–275.
[11] A.W.F. Edwards, Human genetic diversity: Lewontin's fallacy, Bioessays 25 (2003) 798–801.
[12] N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard, Informativeness of genetic markers for inference of ancestry, Am. J. Hum. Genet. 73 (2003) 1402–1422.
[13] L. Ding, H. Wiener, T. Abebe, M. Altaye, R.C. Go, C. Kercsmar, G. Grabowski, L.J. Martin, G.K. Hershey, R. Chakorborty, T.M. Baye, Comparison of measures of marker informativeness for ancestry and admixture mapping, BMC Genomics 12 (2011) 622.
[14] E. Zubritsky, SNP mining. The rush is on, Anal. Chem. 71 (1999) 683A–686A.
[15] S.C. Shah, A. Kusiak, Data mining and genetic algorithm based gene/SNP selection, Artif. Intell. Med. 31 (2004) 183–196.
[16] J. Houle, W. Cadigan, S. Henry, A. Pinnamaneni, S. Lundahl, Database Mining in the Human Genome InitiativeAvailable at: http://www.biodatabases.com/whitepaper.html 2004.
[17] C. Ramu, C. Gemund, CGImodel: CGI programming made easy with Python, Linux J. 75 (2000) 142–149.
[18] J.K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, Genetics 155 (2000) 945–959.
[19] M.B. Kovacic, J.M. Myers, N. Wang, L.J. Martin, M. Lindsey, M.B. Ericksen, H. He, T.L. Patterson, T.M. Baye, D. Torgerson, L.A. Roth, J. Gupta, U. Sivaprasad, A.M. Gibson, A.M. Tsoras, D. Hu, C. Eng, R. Chapela, J.R. Rodriguez-Santana, W. Rodriguez-Cintron, P.C. Avila, K. Beckman, M.A. Seibold, C. Gignoux, S.M. Musaad, W. Chen, E.G. Burchard, G.K. Hershey, Identification of KIF3A as a novel candidate gene for childhood asthma using RNA expression and population allelic frequencies differences, PLoS One 6 (2011) e23714.
[20] T.M. Baye, R.A. Wilke, Mapping genes that predict treatment outcome in admixed populations, Pharmacogenomics J. 10 (2010) 465–477.
[21] R. Moonesinghe, M.J. Khoury, T. Liu, J.P. Ioannidis, Required sample size and nonreplicability thresholds for heterogeneous genetic associations, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 617–622.